


Article

A Deep Learning-Based Parameter Prediction Method for Coal Slime Blending Circulating Fluidized Bed Units

Jiyu Chen , Feng Hong* and Mingming Gao

State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources, School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China; sjyncepu@ncepu.edu.cn (J.C.); gmmncepu@outlook.com (M.G.)

* Correspondence: hongf@ncepu.edu.cn

Featured Application: This research promotes the application of artificial intelligence research in unit operation and monitors unit operation stability using parameter prediction.

Abstract: Coal slime blending can effectively improve the utilization rate of fossil fuels and reduce environmental pollution. However, the combustion in the furnace is unstable due to the empty pump phenomenon during the coal slurry transport. The combustion instability affects the material distribution in the furnace and harms the unit operation. The bed pressure in the circulating fluidized bed unit reflects the amount of material in the furnace. An accurate bed pressure prediction model can reflect the future material quantity in the furnace, which helps adjust the operation of the unit in a timely fashion. Thus, a deep learning-based prediction method for bed pressure is proposed in this paper. The Pearson correlation coefficient with time correction was used to screen the input variables. The Gaussian convolution kernels were used to implement the extraction of inertial delay characteristics of the data. Based on the computational theory of the temporal attention layer, the model was trained using the segmented approach. Ablation experiments verified the innovations of the proposed method. Compared with other models, the mean absolute error of the proposed model reached 0.0443 kPa, 0.0931 kPa, and 0.0345 kPa for the three data sets, respectively, which are better than those of the other models.

Keywords: circulating fluidized bed; bed pressure; deep learning; coal slime; differential prediction



Citation: Chen, J.; Hong, F.; Gao, M. A Deep Learning-Based Parameter Prediction Method for Coal Slime Blending Circulating Fluidized Bed Units. *Appl. Sci.* **2022**, *12*, 6652. <https://doi.org/10.3390/app12136652>

Academic Editor: Abílio M. P. De Jesus

Received: 23 May 2022

Accepted: 28 June 2022

Published: 30 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

China is advocating coal washing and quality improvement processing as well as promoting coal energy low-carbon development. Coal slime is a byproduct of the coal washing process and is characterized by fine particle size, high water holding capacity, high viscosity, high ash content, and low heat generation [1]. If the coal slime is not correctly disposed of, it will harm the environment. The harmful substances contained in coal slime can seep into the ground and cause soil contamination. If coal slime is discharged into rivers and lakes, it will cause severe pollution to water quality. Circulating fluidized bed combustion (CFB) technology, one of the clean coal technologies, has strong adaptability to medium-quality and low coal fuel, with combustion efficiency reaching 95~99% [2]. Therefore, CFB coal slime blending is recognized as the best way to handle coal slime.

However, coal slime blending affects the stability of CFB units significantly. Ideally, the water content of coal slime is constant in each pump of the coal slime pump. In actual operation, the coal slime falls and splashes water on the level meter, causing the alarm. The alarm causes the discharge screw to stop working, emptying the pump. As a result, the uneven mixture of coal slime and water content significantly fluctuates in boiler input energy. The bed pressure of the coal slime blending unit shows a wide range of fluctuations due to the unevenness of the coal feed in the furnace. In the pant-leg CFB units, the

unstable bed pressure will lead to the inventory overturn accident, threatening the stable operation of the unit. Accurate prediction models improve the safety of unit operation. By obtaining future information through forecasting, operators can make timely adjustments to their operations, thereby reducing accidents. Therefore, it is necessary to monitor and predict bed pressure. Most studies in bed pressure only focus on mechanism research. The experimental relationship between the lateral pressure difference of the upper furnace and the lateral material exchange flow rate was established by Li [3]. Li also found the critical role of the lateral pressure difference of the upper furnace in material fluctuation, which was also verified by Sun [4]. Many researchers focused on the solid flow model in the furnace to study bed pressure [5]. Yang [6] developed a computational particle fluid dynamics model to simulate solid exchange behavior between two half beds in the bench-scale two-dimensional dual-leg fluidized bed. Wang [7] proposed an empirical correlation of the non-uniformity index to quantify the non-uniformity of gas–solid two-phase flow. The computational particle fluid dynamics model of a pilot-scale CFB was used to numerically simulate its gas–solid flow characteristics by Liu [8]. Afsin Gungor [9] established a CFB axial pressure distribution prediction model based on the particle method to predict bed pressure. Current research also shows that the airflow rate in the furnace has a significant influence on the bed pressure [10]. However, the structure of these mechanism models is complex, and it is difficult to obtain some of the data in the mechanism models. In addition, the empty pump phenomenon makes the measured value of feed quantity in the unit deviate from the actual value. Therefore, the mechanism models make it challenging to achieve accurate bed pressure prediction in coal slime blending CFB units.

Driven by the development of intelligent algorithms, data-driven modeling technology for complex industrial processes is attracting the attention of researchers [11,12]. The data-driven models are applied and migrated without considering the design parameters of units. Such models only need to use operational data to adjust and train the parameters of the model. The error backpropagation neural network (BPNN) [13], least squares support vector machine (LSSVM) [14], and other algorithms are commonly used in the industrial process [15,16]. Due to the time relevance character, algorithms that focus on analyzing the relationships between time relevance industrial data are more suitable for typical industrial process modeling. The emergence of the recurrent neural network (RNN) and the long short-term memory neural network (LSTM) improves the ability of neural network models to extract temporal information from temporal data. These two neural networks are also widely used in industrial forecasting. Xia [17] proposed a new method for predicting renewable energy generation and power load in univariate and multivariate scenarios using the improved stacked gated recurrent unit neural network. The method uses various weather parameters to predict wind power generation and historical energy consumption data to predict power load. Shahid [18] proposed a model composed of LSTM and the genetic algorithm to predict short-term wind power generation. It also provides accurate, reliable, and robust predictions for the wind power of seven wind farms in Europe. Inapakurthi [19] used RNN and LSTM to capture the dynamic trends of 15 environmental parameters, including particulate matter and pollutants in the atmosphere that cause long-term health hazards. In addition to these models, the attention models have also received attention from researchers. The idea of Attention originated from visual images in the 1990s. In 2014, Google Mind [20] team used the attention mechanism to classify images on the RNN model and achieved excellent results. Attentional mechanisms are used in many areas, such as the machine translation field [21], the image recognition field [22], and the language emotion analysis field [23]. In recent years, attention was also widely used in the prediction field. Azam [24] proposed a new hybrid deep learning method based on bi-directional long short-term memory and a multi-head self-attention mechanism, which can accurately predict the marginal price of position and the system load one day ago. By analyzing the formation mechanism of NO_x and the reaction mechanism of the SCR reactor, a sequence-to-sequence dynamic prediction model was proposed by Xie [25], which can fit multivariable coupling, nonlinear, and large delay systems. Shih [26]

proposed a novel attention mechanism for selecting the relevant time series. Shih used its frequency domain information for multivariate forecasting.

Influenced by the successful application of temporal pattern attention [26], an improved attention layer is proposed in this paper. The model uses the Gaussian convolution kernels in the convolutional neural network, enabling the model to extract different inertial delay properties. Scaled dot-product attention [27] is used in the model for attention computation. The prediction method in this paper treats the input and output of the model and improves the way the model is trained. The main contributions of this paper are summarized as follows:

(1) The delayed relationship between variables was considered in the correlation analysis. The best combination of variable inputs was selected through comparison experiments.

(2) The differential values of the input variables are used as additional input features, which reduce the effect of the autocorrelation of the inputs on the model. The differential values in bed pressure are used as the predictive target for the model to improve the learning ability of the model for the target value.

(3) The attention layer uses the Gaussian convolution kernel to extract the inertial and delay properties of the features. The convolution operation allows the model to learn different inertial and delay properties for different features. Thus, the prediction performance of the model is improved.

(4) Based on the principle of the attention mechanism, the model is trained in segments. The first training segment enables the query vector in the attention layer to learn the target information well. The parameters of the attention layer are updated in the second segment training.

The structure of this paper is as follows: Section 2 describes the background and principle of the method used in this paper. In Section 3, the forecast method proposed in this paper is used to forecast the actual operation data. In Section 4, the prediction methods proposed in this paper are used for ablation research and comparison of prediction models. The ablation study verified the effectiveness of the model structure and data processing method in the prediction method. The model proposed in this paper performs better than other algorithm models in comparing forecast models. Finally, Section 5 summarizes the conclusions obtained from the study with the highlights of significant findings.

2. Background Description

2.1. Coal Slime Blending Circulating Fluidised Bed Overview and Differential Prediction

Alstom pioneered a pant-leg circulating fluidized bed design for its 185 MW unit in New Mexico [28]. The pant-leg boiler structure enhances the secondary air penetration capability. This structure overcomes the difficulties of larger circulating fluidized bed units [29]. The pant-leg circulating fluidized bed boiler structure is shown in Figure 1. The coal and limestone are crushed to proper particle size and mixed with hot boiling materials in the furnace. Among them, the size range of pulverized coal particles is 0~12 mm, $d_{50} = 1.1$ mm, and the size range of limestone particles is 0~1 mm, $d_{50} = 0.3$ mm. The coal slime is pumped into the furnace chamber to participate in the combustion. The furnace temperature is usually controlled at 850–920 °C to promote the desulfurization reaction. Under the elutriation phenomenon of the CFB unit, part of the small particle will be carried out of the furnace with the flue gas, separated by the cyclone separator, and then returned to the furnace to continue to engage in combustion [30].

Pumping is the most widely used method of blending coal slime in large proportions in CFB boilers. The process of coal slime conveying is shown in Figure 2. The parts with green represent the normal conveying process. The red parts indicate the reasons for the empty pump phenomenon in actual operation. Under normal conditions, the coal slime stored in the coal slime bin is fed to the mixing hopper via the bin bottom slide frame and discharge screw. The coal slime is then pressurized by the plunger coal slime pump and sent to the coal slime lance for injection into the furnace. However, in actual operation, the coal slime mixing is mainly carried out in the hopper. The water enters the hopper

and the conveying process all the time. Therefore, the water of the coal slime into the furnace fluctuates greatly. The water of the incoming coal slime fluctuates considerably. The coal slime often falls and splashes water onto the level meter. The water on the level meter causes the level meter to display inaccurately. The safety system sends an alarm signal, stopping the discharge screw from working. In this case, no coal slime enters the mixing hopper, but the plunger coal slime pumps and the moisture regulation system are still running, resulting in the empty pump phenomenon. The fluctuating moisture content of the coal slime and the empty pump phenomenon leads to unstable combustion in the furnace. The combustion instability ultimately leads to an unstable material distribution in the furnace.

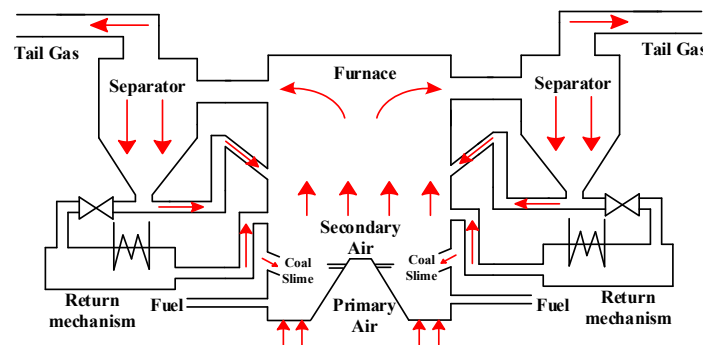


Figure 1. Structure of pant-leg circulating fluidized bed boiler.

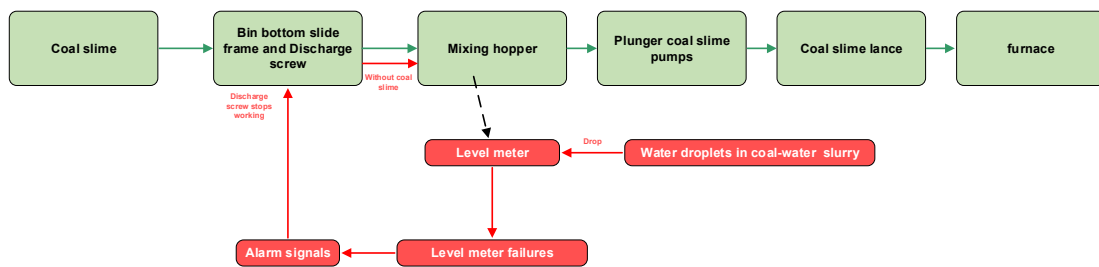


Figure 2. Coal slime conveying flow chart.

The materials on both sides of the furnace chamber are exchanged through the dilute phase area to ensure the material balance during regular operation. When there is a significant deviation in the air quantity on both sides, the side with the larger air quantity will blow the materials to the other side. Due to the accumulation of materials, the air quantity on the other side is reduced. In this case, the side with a large air quantity continuously transfers the material to the other side, forming positive feedback and leading to inventory overturn accidents. The natural phenomenon of inventory overturn accidents is the serious deviation of bed pressure on both sides of the boiler. The process of the inventory overturn of the CFB unit is shown in Figure 3.

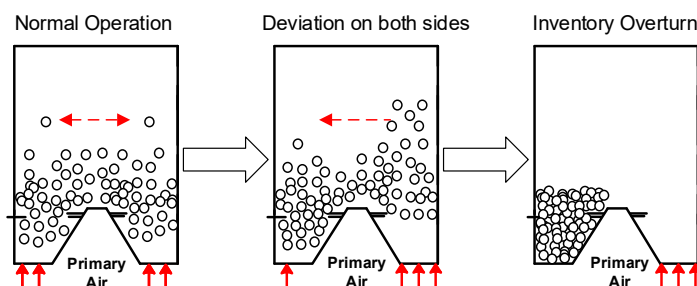


Figure 3. The process of inventory overturn. The solid arrows represent the direction of the airflow. The dashed arrows represent the direction of material flow.

The bed pressure of the CFB unit is mainly affected by the amount of material in the furnace [3]. From this perspective, the bed material quantity can be divided into the initial material quantity and the variation of material quantity. The variation of material quantity is caused by the change of unit state, which is usually affected by a particular data segment of the unit. The initial material quantity is determined by the whole operation data of the unit. The model input is a time series of data in the prediction task. Therefore, it is difficult for the model to learn the initial material quantity by direct prediction.

In the bed pressure prediction task, the predicted value is assumed to be the bed pressure value y_{t+k} at time $t + k$, and the input to the model is the data at time t and before. Therefore, we assume that the predicted value y_{t+k} is composed of the bed pressure variation value Δy and the initial bed pressure value y_t . The formula is shown in Equation (1). The Δy is predicted by the prediction model. The y_t is obtained from the current bed pressure measuring point. The method in this paper calculates y_{t+k} by predicting Δy .

$$y_{t+k} = \Delta y + y_t \tag{1}$$

where t is the current moment; k is the prediction step.

Because of the complex combustion process of CFB units, much literature uses the first-order differential equation to model [31,32]. Inspired by these studies, this paper takes the differential value of variables as an additional input feature. In addition, the differential treatment reduces the effect of the autocorrelation of the input variables on the model. The input difference score formula is shown below.

$$\Delta x_t = x_t - x_{t-k} \tag{2}$$

where x_t denotes the model input at time t ; Δx_t denotes the additional input features of the model.

2.2. Long Short-Term Memory

The long short-term memory network (LSTM) was proposed by Schmidhuber and Hochreiter in 1997 [33]. The data were put through several such processes in the network. The operation process is as follows:

$$i_t = \sigma(W_i \cdot [x_t, h_{t-1}] + b_i) \tag{3}$$

$$f_t = \sigma(W_f \cdot [x_t, h_{t-1}] + b_f) \tag{4}$$

$$o_t = \sigma(W_o \cdot [x_t, h_{t-1}] + b_o) \tag{5}$$

$$\tilde{c}_t = \tanh(W_c \cdot [x_t, h_{t-1}] + b_c) \tag{6}$$

$$c_t = f_t \times c_{t-1} + i_t \times \tilde{c}_t \tag{7}$$

$$h_t = o_t \times \tanh(c_t) \tag{8}$$

where W and U represent weight values; b represents offset values; h_{t-1} represents the output of its cells at the previous moment; f , i , o denote the forgetting gate, the input gate, and the output gate, respectively; c represents the cell state of the neuron; \tilde{c} indicates the candidate cell state of the neuron; h_t denotes the output of the LSTM layer at time t ; and the subscript t indicates the moment t .

σ represents the sigmoid activation function. The expression of the sigmoid activation function and the tanh activation function are as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{9}$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{10}$$

2.3. Correlation Coefficient Calculation

There are different delay characteristics between the sampling data of different characteristics of CFB boilers. Such delay characteristics have an impact on the similarity analysis between variables. The traditional Pearson correlation coefficient does not consider the influence of delay characteristics between variables on the correlation coefficient. In this paper, the time correspondence between the two sequences is corrected by sliding calculation. Moreover, the Pearson correlation coefficient of the corresponding time correction value is calculated. The maximum value of all correlation coefficients is used as the final correlation coefficient. The algorithmic process is as follows:

(1) Randomly generated $t \subseteq N_+$, $d_{max} + l - 1 \leq t \leq L_{train} - k$. L_{train} is the size of the training set, and l is the length of the data sequence. d is the time correction value, and d_{max} is the maximum of d .

(2) Data sequence $X\{x_{t-d-l+1}, x_{t-d-l+2}, \dots, x_{t-d}\}$ and $Y\{y_{t-l+k+1}, y_{t-l+k+2}, \dots, y_{t+k}\}$ are sampled in the input and output of the training set.

(3) d changes from 0 to $d_{max} - 1$. The Pearson correlation coefficients of sequence $X\{x_{t-d-l+1}, x_{t-d-l+2}, \dots, x_{t-d}\}$ and $Y\{y_{t-l+k+1}, y_{t-l+k+2}, \dots, y_{t+k}\}$ are calculated, respectively. A set of correlation coefficients with the length of d_{max} is obtained. The Pearson correlation coefficient formula is shown in Equation (11).

$$\rho_{d,X,Y} = \frac{\sum_{i=t-d-l+1}^{t-d} (x_i - \bar{x}) \times (y_{i+k+d} - \bar{y})}{\sqrt{\sum_{i=t-d-l+1}^{t-d} (x_i - \bar{x})^2 \times \sum_{i=t-d-l+1}^{t-d} (y_{i+k+d} - \bar{y})^2}} \tag{11}$$

where \bar{x} and \bar{y} represent the average of the sampling sequences X, Y , respectively; x_i and y_i represent the i -th x in the X , and the i -th y in the Y , respectively.

(4) Repeat step 1 to step 3 for M_{re} times. The average value $\bar{\rho}_{d,X,Y}$ of M_{re} results is calculated.

(5) The maximum value in $\bar{\rho}_{d,X,Y}$ is used as the similarity $\bar{\rho}_{X,Y}$ between variable X and Y .

$$\bar{\rho}_{X,Y} = \max_d (\bar{\rho}_{d,X,Y}) \tag{12}$$

2.4. Attention Layer

Inspired by the successful application of temporary pattern attention in the literature [26], this paper proposes an improved attention network structure. Figure 4 shows the calculation process of the model. The model extracts temporal information H_{t-1}^C in H_{t-1} by the convolution operations. Where $H_{t-1} = \{h_1, h_2 \dots, h_{t-1}\}$, h_t represents the LSTM layer output at time t . By focusing on the output h_t at the last moment, the model extracts the relevant time sequence information in H_{t-1}^C , thereby improving the learning ability of the model. In convolution operation, the Gaussian convolution kernel is used to extract the inertial delay characteristics of data. The convolution kernels corresponding to each feature are independent of each other. Scaled dot-product attention is used to calculate the similarity of time series. The Gaussian convolution kernel formula used in the model is as follows:

$$ker = K_{ker} \times \exp\left(-\frac{(a-b)^2}{c^2}\right) \tag{13}$$

$$a_i = i \tag{14}$$

$$K_{ker} = \begin{cases} 1, & a-b \geq 0 \\ 0, & a-b < 0 \end{cases} \tag{15}$$

where $a, b, c \in \mathbb{R}^w$, w is the length of the convolution kernel; a_i representing the i -th element in a ; b , and c are trainable parameters. The shape of the Gaussian convolution kernel depends on b and c . Different Gaussian convolution kernels can extract different inertia and

delay characteristics in the data. K_{ker} is a step function used to cut the Gaussian convolution kernel. It should be noted that in model training, K_{ker} does not perform gradient calculation.

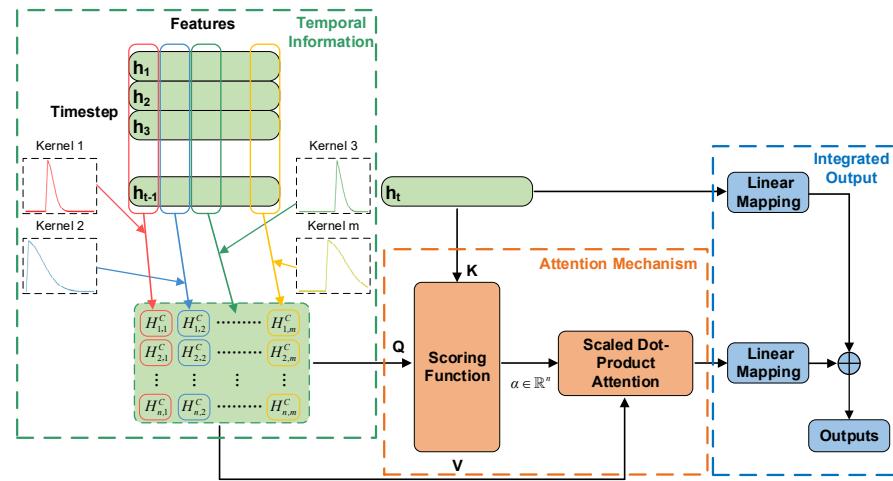


Figure 4. Attention layer structure diagram. Where $n = t - w$, t is the length of time step of input data; w is the size of the convolution kernel; and m is the characteristic number of input data.

The convolution operation can extract the data delay inertia characteristics through the Gaussian convolution kernel. To better illustrate, we randomly generate a Gaussian convolution kernel for elaboration. We assume that the data has an input of 1 at the moment 0 and 0 at other moments. We pad the input at the 0 moment with $w - 1$ zeros in front of it. The result of the convolution operation is the influence of the input at the moment 0 on the input at subsequent moments. The result of the convolution operation is shown in Figure 5. The data first pass through a purely delayed segment and then gradually makes an impact, reaching a peak and then stopping. This inertial delay response is close to the step response curve of each variable and the target variable in Reference [34]. The trainable parameters in the convolution kernel determine the shape of the convolution kernel so that the model can extract the inertial delay characteristics of different characteristics.

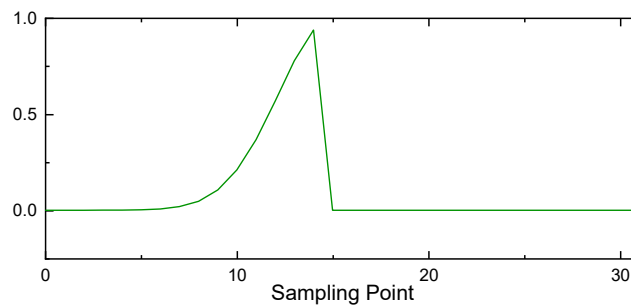


Figure 5. Data impact diagram.

The calculation process of the model is as follows: The model first performs the convolution operation on the data, and H_{t-1} forms H_{t-1}^C after the convolution operation.

$$H_{t-1}^C = CNN(H_{t-1}) \tag{16}$$

The similarity between h_t and the feature at each time step in H_{t-1}^C is calculated. The attention calculation method is scaled dot-product attention, with H_{t-1}^C as the query vector and value vector, and h_t as the key vector. v_{t-1} is obtained by weighted summation of H_{t-1}^C according to the similarity score. v_{t-1} and h_t are weighted and summed to obtain the final output O .

$$\alpha = softmax\left(f\left(H_{t-1}^C, h_t\right)\right) \tag{17}$$

$$f(H_{t-1}^C, h_t) = H_{t-1}^C h_t / \sqrt{m} \quad (18)$$

$$v_{t-1} = \alpha H_{t-1}^C \quad (19)$$

$$O = W_h \times h_t + W_v \times v_{t-1} \quad (20)$$

where $H_{t-1} \in \mathbb{R}^{(t-1) \times m}$, $H_{t-1}^C \in \mathbb{R}^{n \times m}$, $W_h \in \mathbb{R}^{m \times m}$, $W_v \in \mathbb{R}^{m \times m}$, $v_{t-1} \in \mathbb{R}^m$, $\alpha \in \mathbb{R}^n$.

3. Structuring of the Proposed Forecast Framework

In this section, the production data from LinHuan Zhongli 1 # 330 MW circulating fluidized bed unit in China illustrates the prediction modeling steps. The boiler slime blending system was designed and manufactured by PUTZMEISTER, Germany. The coal slime blending method is a typical coal slime pumping system consisting of the coal slime silo, the bin bottom slide frame, the flushing water pressure pump, and four coal slime pumping pipelines. The coal slime guns are arranged at a position 2.6 m high from the air distribution board, two on each of the left and right side walls of the boiler, and arranged horizontally and symmetrically. The coal slime gun comprises the gun body, the ball valve, the gate valve, and the safety valve. This paper uses three datasets for prediction and comparison. Among them, dataset 1 and dataset 2 come both from the operational data from 18 to 20 June 2018. The difference is that dataset 1 is all the operational data from the unit during regular operation. Dataset 2 has fault data in the test set. Dataset 3 is the operating data of the unit between 12 September and 14 September 2015. All datasets have a sampling interval of 5 s and a prediction step of 6. The size of all datasets is 39,600. All data sets are divided into training, validation, and test sets according to the data segment lengths of 36,000, 1600, and 2000 in that order. The prediction task is shown as an example of left bed pressure prediction. The three datasets are described as follows:

- Dataset 1: The entire dataset is in regular operation. The time range of the data set was from 18 June 2018, 0:21 to 20 June 2018, 7:21.
- Dataset 2: There was an inventory overturn accident in the test set. The time range of the data set was from 18 June 2018, 14:05, to 20 June 2018, 21:05.
- Dataset 3: The data in the dataset are normal operation data. The time range of the data set was from 12 September 2015, 0:21 to 14 September 2015, 7:21.

The flow of the prediction method in this paper is shown in Figure 6. Among them, the variable screening method is described in Section 2.3. The hyperparameters of the model are adjusted according to the errors on the validation set. The program was compiled using Python, and the algorithm model used Pytorch and Scikit-learn framework. All experiments were carried out in the Python compiling environment using an Intel Core i9-10900K CPU and RTX3080 GPU machine. CUDA and cuDNN accelerate the algorithm model.

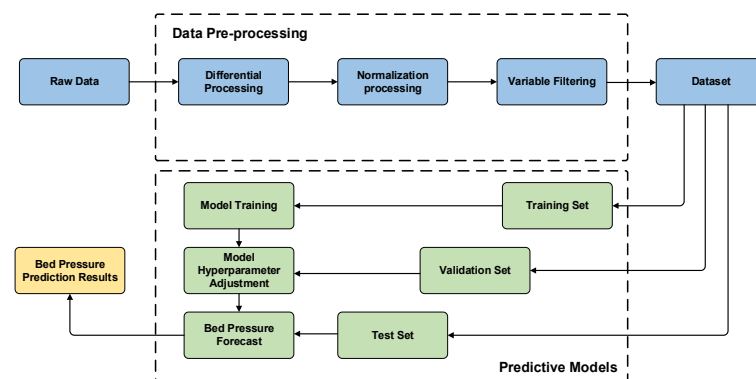


Figure 6. Prediction flow diagram. The data are sequentially differenced and normalized. The variables are filtered by training set and validation set. Finally, the parameters of the model are trained by the training set. The hyperparameters of the model are selected by the validation set.

3.1. Prediction Model Structure

The prediction model structure proposed in this paper is shown in Figure 7. The variable data is input into the network model after the variable screening. The number of input features is $m/2$. The input data are differenced to form the new input to the model, where the differential values of the model inputs are also normalized. The step length of the differential value is the same as the length of the prediction step. The final model input with feature number m is formed. In this model, the LSTM part is composed of a stack of N identical layers. Each layer has two sub-layers. The first is an LSTM layer with $2m$ units, and the second LSTM layer with m units. The attention model in this paper computes the data after passing through the LSTM part. The final output is formed by the dense layer, the batch normalization layer, and the activation function. The output of the model is the differential value between the current time and the predicted time.

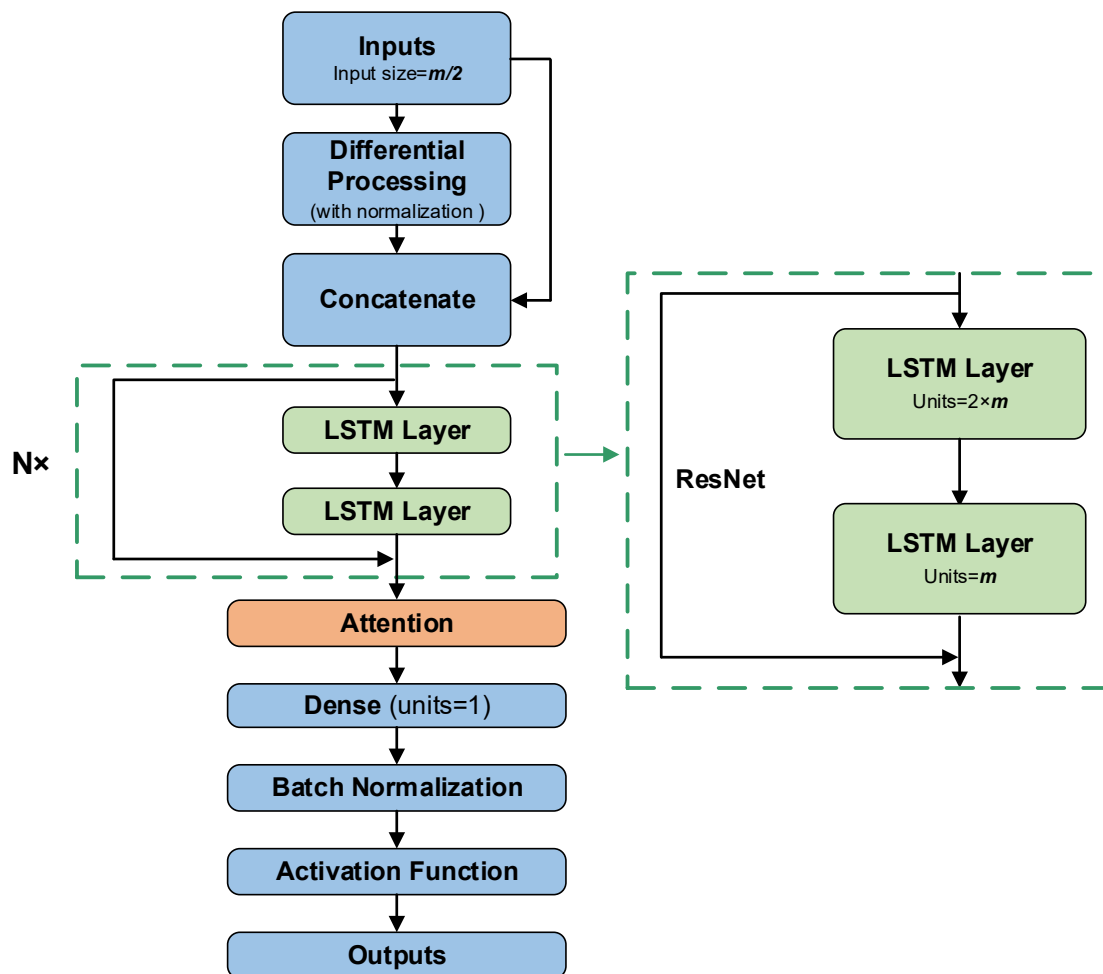


Figure 7. The structure of the prediction model.

The dense layer adopted the full connection layer of one neuron in the model structure, and the activation function adopted the Sigmoid function. The function expression is shown in Equation (9). The ResNet (residual network) structure and batch normalization layer were added to the model structure in this paper. The ResNet and batch normalization are described as follows:

- ResNet: This network structure was proposed in 2015 [35], which significantly promoted deep learning model development. This structure is added to solve the problem of model degradation in the neural network, which refers to the fact that the perfor-

mance drops rapidly after adding more layers to the network. With the addition of ResNet, the parameters of the deep neural network can be optimized more easily.

- **Batch Normalization:** This network layer structure was proposed in 2015 to solve interval covariate shift (ICS) [36]. It is found that the layer has the advantage of smoothing the optimization space and regularizing the model in the subsequent research. Therefore, the batch normalization layer is used before the final output layer in the model to accelerate the convergence of the model.

3.2. Segmented Training

In this work, a segmented training approach was applied to learn the parameters of the model based on the computation of the attention layer. In the original temporal pattern attention calculation process, the attention layer in this paper focuses on the last-moment state output h_t of the LSTM layer. In the early stages of training, slow convergence of the attention layer parameter learning occurs due to the information confusion in h_t . Therefore, this paper uses a segmented training approach to train the parameters of the model. The model is divided into an LSTM layer part and an output layer part, using the attention layer as the dividing point. A fully connected layer with one unit is added in the first segment after the h_t . A sigmoid activation function is used for the final output. After the training, the fully connected layer is replaced with the output layer part of this paper. In the second segment of training, the parameters of the LSTM layer part are fixed, and only the parameters of the output layer part are trained. The training process is shown in Figure 8.

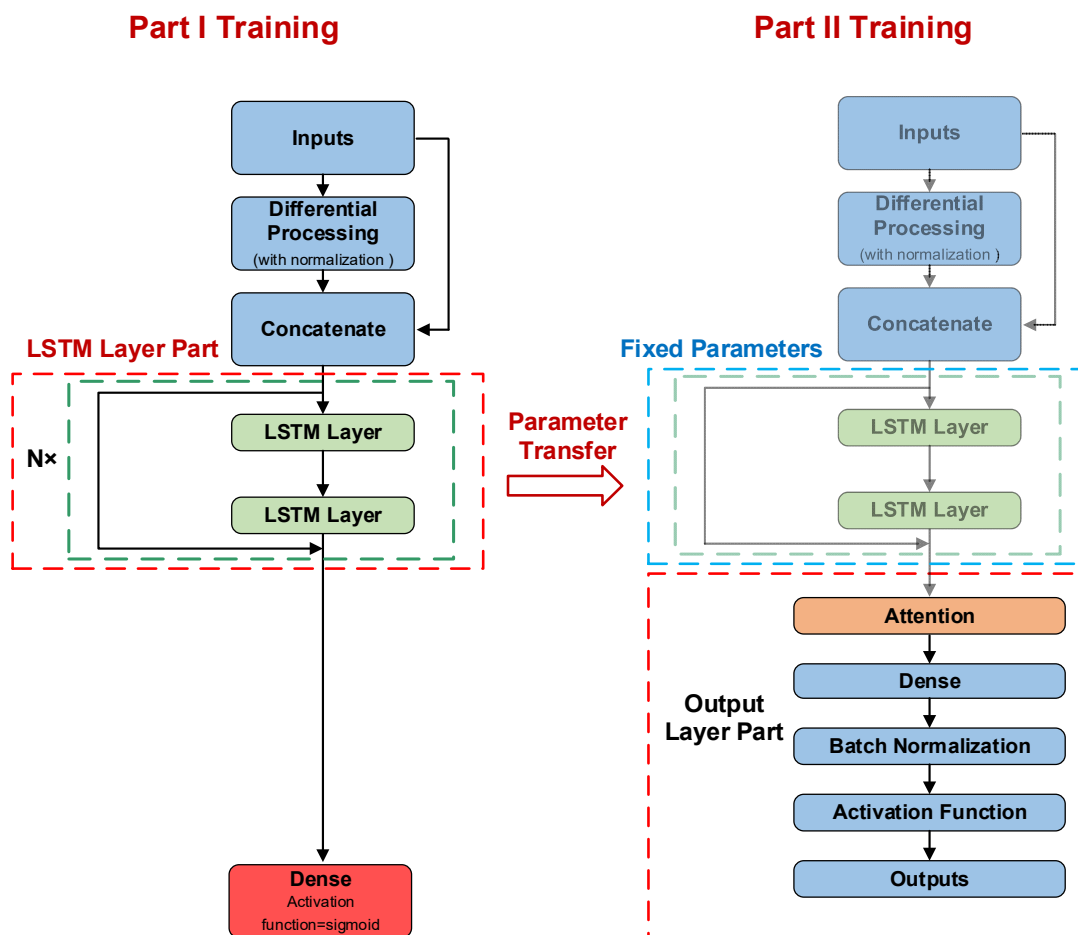


Figure 8. Segmented training diagram. The red box in the figure indicates the position of parameter updates in each training segment. The blue box indicates the positions where the parameters are fixed.

3.3. Performance Assessment

To assess the prediction performance under different experimental scenarios, two scientific performance metrics are selected for prediction. This paper chooses mean absolute error (MAE) and mean absolute percentage error (MAPE) as performance metrics, which are used for evaluating the performance of different models in prediction results and can be expressed as follows:

Mean absolute error (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - y'_i| \quad (21)$$

Mean absolute percentage error (MAPE):

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - y'_i}{y_i} \right| \quad (22)$$

where y'_i is the predicted value; y_i is the target value; n is the number of prediction data; \bar{y} is the average value of y_i . Generally, the lower values of MAE and MAPE lead to better performance of the prediction task.

3.4. Data Standardization and Differential Processing

Too large or too small an input can easily lead to unstable gradient values. Data normalization can effectively avoid such problems. This paper uses the min-max scaling method to normalize the data linearly. The min-max scaling formula is as follows: mapping the data to the range of 0 to 1 by min-max scaling is more beneficial to the training of network model parameters.

$$x'_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (23)$$

where x_{\min} and x_{\max} represent the minimum and maximum values in the data; x'_i represents the data after standardization; x_i represents the data before standardization.

The prediction method in this paper uses a differential prediction method, which is stated in Section 2.1. The differential processing is divided into differential processing for the target value and differential processing for the input. The differential processing for the target value is shown below.

$$\Delta y = y_{t+k} - y_t \quad (24)$$

where y_{t+k} is the value of the predicted target at time $t + k$; y_t is the value of the predicted target at time t ; Δy is the differential value between the two moments. The prediction model obtains the value of y_{t+k} by predicting Δy . The k is set to 6 in this paper.

In addition, the prediction method in this paper also performs differential processing on the input of the model. The differential processing for the input is shown in Equation (2).

3.5. Variable Filtering

The method described in this paper uses Dataset 1 to filter the variables. CFB unit control quantities are used as input features for the model in this paper. The differential values of these variables are input as features of the model. The correlation coefficients between the input features and differential value were calculated using the method in Section 2.3. This paper sets d_{max} , l , and M_{re} to 100, 200, 4000, respectively. Table 1 shows the calculation results of the correlation coefficients. The correlation coefficients of the input variables are sorted in descending order, and the results are shown in Table 1.

The screening of variables based on empirical thresholds is unreliable, so the comparison experiments are used to perform variable screening. The experiments first rank the importance of the variable features according to the correlation coefficient. Then the experimental groups were constructed by adding the features in order of importance. Finally, the accuracy performance of the prediction model on the validation set is used as the basis for

variable selection. Table 2 shows the prediction performance of each experimental group on the validation set. EG-i indicates that the top i most essential features among the features are used as the input to the model. The same model hyperparameter settings were used for all experimental groups, as shown in Table 3. Table 2 presents the mean and standard deviation of all experimental groups in three runs. The best performance is in boldface.

Table 1. The results table of correlation coefficient.

Variable Name	Correlation Coefficient	Rank Value
Fuel quantity	0.33070	1
Rotation speed of left slag cooler	0.16932	9
Rotation speed of right slag cooler	0.14856	11
Valve opening of left return valve	0.20790	8
Valve opening of right return valve	0.14490	12
Secondary air quantity at lower left side	0.27176	7
Secondary air quantity at lower right side	0.28662	5
Secondary air quantity at upper left side	0.29472	3
Secondary air quantity at upper right side	0.27559	6
Primary air quantity on the left side	0.28987	4
Primary air quantity on the right side	0.31606	2
Left limestone flow	0.15164	10
Right limestone flow	0.08738	13

Table 2. The experimental results table of variable selection.

Group Name	MAE	MAPE/%
EG-1	0.0570 ± 0.0001	10.47 ± 0.03
EG-2	0.0491 ± 0.0002	9.02 ± 0.15
EG-3	0.0461 ± 0.0004	8.40 ± 0.06
EG-4	0.0433 ± 0.0010	7.82 ± 0.14
EG-5	0.0425 ± 0.0005	7.74 ± 0.18
EG-6	0.0404 ± 0.0010	7.44 ± 0.23
EG-7	0.0446 ± 0.0002	8.21 ± 0.08
EG-8	0.0445 ± 0.0003	8.14 ± 0.06
EG-9	0.0448 ± 0.0004	8.03 ± 0.04
EG-10	0.0456 ± 0.0017	8.12 ± 0.18
EG-11	0.0489 ± 0.0019	8.63 ± 0.29
EG-12	0.0461 ± 0.0018	8.25 ± 0.37
EG-13	0.0466 ± 0.0008	8.26 ± 0.12

Table 3. The hyperparameter table of the variable screening model.

Hyper-Parameters	Title 3
Timestep	64
N	1
Batch size	100
Learning rate	0.001
Size of the convolutional kernel	8

As can be seen from Table 2, the best prediction accuracy of the validation set was achieved for the input features of EG-6. Therefore, the input features of EG-6 are considered the results of the variable filtering. The selected variables are fuel quantity, primary air quantity on the right side, secondary air quantity at upper left side, primary air quantity on the left side, secondary air quantity at lower right side, and secondary air quantity at upper right side.

3.6. Model Prediction Results

The model is trained using the segmented training approach. The parameters of the LSTM layer are trained first. The last moment output h_t of the LSTM layer is output through a fully connected layer. The sigmoid activation function is used for the final output. After the first training segment, the model parameters of the LSTM are migrated into the model of this paper. In the second training segment, only the parameters of the output layer part are trained. The hyperparameters of the model in this paper are selected by the validation set. The specific hyperparameters are shown in Table 4.

Table 4. The hyperparameter selection results of prediction model.

Hyper-Parameters	Title 3
Timestep	128
N	2
Batch size	50
Learning rate	0.001
Size of the convolutional kernel	32

The model was trained on 36,000 training samples. The average training time in the first training segment was 650 s, and the average period of training stop was 134. The average training time in the second training segment was 260 s, and the average period of training stop was 83. The entire training process is performed using the early stopping strategy, and the optimizer uses AdamW [37]. Figures 9 and 10, and Table 5 show the predicted results of the model on the test set. The mean values of the predicted results are presented in Figures 9 and 10. The mean and standard deviation of the results of the three replicate experiments are shown in Table 5. In Figure 9, the target value is the actual differential value. In Figure 10, the target value is the actual value of the bed pressure after 30 s. From the results, it can be seen that the model in this paper has accurate prediction performance on all three datasets. It can be seen from the prediction results of Dataset 2 that the model can still capture the bed pressure changes well in the case of sudden changes in bed pressure.

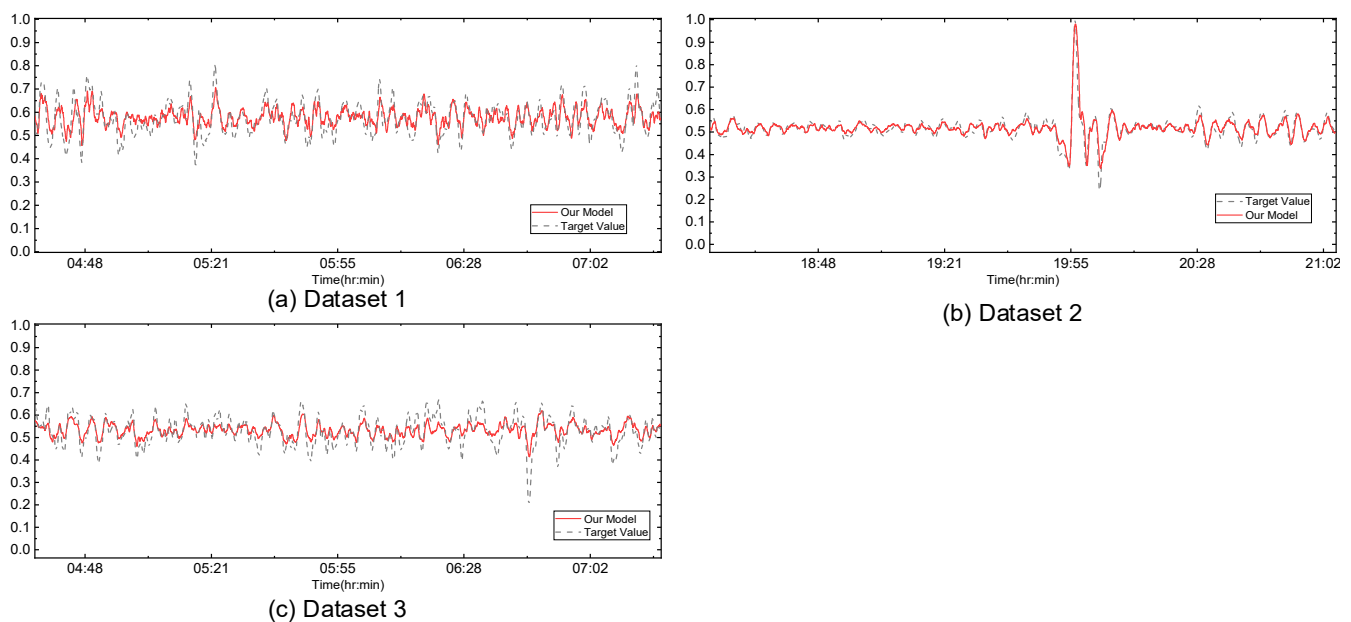


Figure 9. The prediction results diagram of differential value.

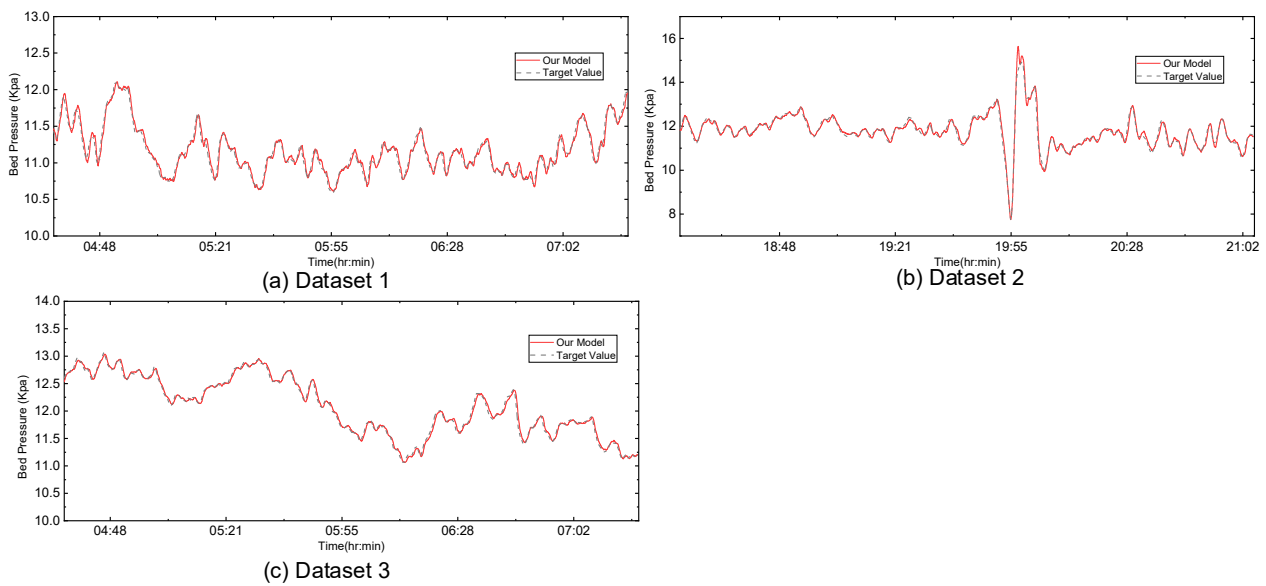


Figure 10. The prediction results diagram of bed pressure.

Table 5. Prediction results table.

	Bed Pressure Value		Differential Value	
	MAE/kPa	MAPE/%	MAE	MAPE/%
Dataset 1	0.0443 ± 0.0004	0.40 ± 0.00	0.0379 ± 0.0003	6.81 ± 0.07
Dataset 2	0.0931 ± 0.0030	0.79 ± 0.03	0.0165 ± 0.0005	3.33 ± 0.12
Dataset 3	0.0345 ± 0.0022	0.29 ± 0.02	0.0327 ± 0.0021	6.68 ± 0.39

4. Comparison and Discussion

In this section, the model proposed in this paper is analyzed and discussed in two main parts. The first part is the ablation study, where several innovations of the prediction method in this paper are ablated for experiments to verify the effectiveness of the innovations. The second section compares the model in this paper with other prediction models. It should be noted that the curves in the pictures in this section are the average of all replicate experiments. In order to show the comparison effect more clearly, the pictures only show the part of the test set with a length of 600. The table shows the mean and standard deviation of the error of all replicate experiments.

4.1. Ablation Study

4.1.1. Differential Prediction Method

The ablation experiments are conducted for the differential prediction method. The differential prediction method was used for the control group. The experimental group used the direct prediction method. Because of the change in the prediction method, the final bed pressure prediction results are used as the target value. The results are shown in Table 6 and Figure 11.

Table 6. The prediction comparison table of differential prediction method.

	Control Group		Experimental Group	
	MAE/kPa	MAPE/%	MAE/kPa	MAPE/%
Dataset 1	0.0443 ± 0.0004	0.40 ± 0.00	0.3779 ± 0.0399	3.39 ± 0.36
Dataset 2	0.0931 ± 0.0030	0.79 ± 0.03	0.4985 ± 0.0404	4.22 ± 0.32
Dataset 3	0.0345 ± 0.0022	0.29 ± 0.02	0.5816 ± 0.0771	4.91 ± 0.71

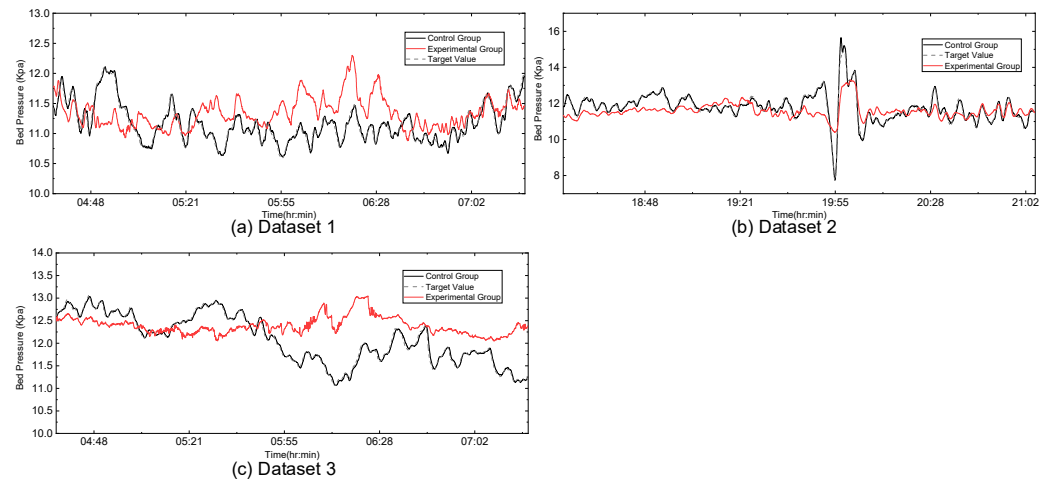


Figure 11. The prediction comparison diagram of the differential prediction method.

It can be seen from the results that the use of the differential prediction method greatly improves the prediction performance of the model. The direct prediction approach makes it difficult for the model to learn the mapping relationship between the input and the bed pressure. The main reason is that the model has a limited input length and cannot fully characterize the initial values of bed pressure resulting from the entire past operation. After adopting the differential prediction approach, the model predicts by learning the bed pressure variation, thus achieving better prediction performance. Moreover, the neural network model is more likely to learn low-frequency information and ignore high-frequency information during the training process [38]. The differential processing of bed pressure similarly extracts the high-frequency information of the data, allowing the model to learn the high-frequency information of the variables directly. Therefore, the predictive performance of the model is improved.

4.1.2. Differential Processing of Input Variables

Ablation experiments are performed for the differential processing of model inputs. Figure 12 and Table 7 show the comparison results of the ablation experiments. The model adding the differential features can learn the differential value features better from the results. The main reason for this is that the differential treatment of the input variables conforms to the form of a first-order differential equation, which improves the learning ability of the model for the unit data. Moreover, the differential treatment reduces the effect of the autocorrelation of the input variables on the model.

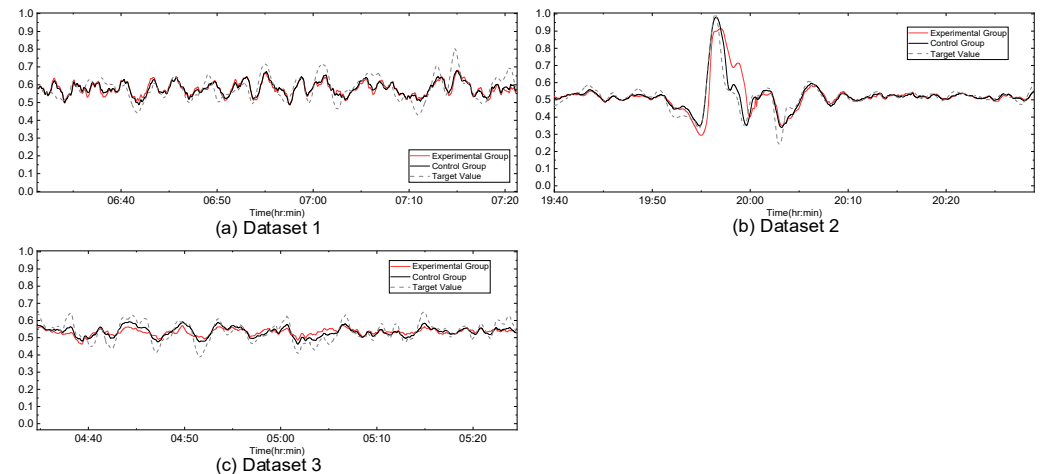


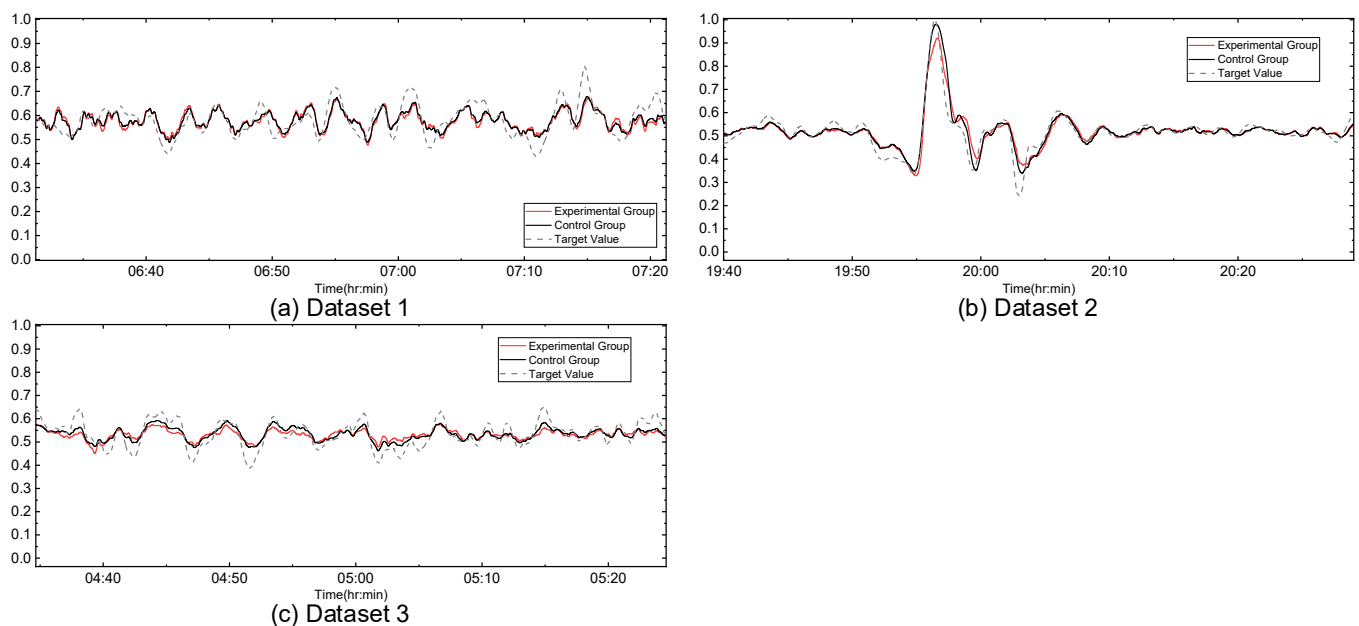
Figure 12. The prediction comparison diagram of differential processing of input variables.

Table 7. The prediction comparison table of differential processing of input variables.

	Control Group		Experimental Group	
	MAE	MAPE/%	MAE	MAPE/%
Dataset 1	0.0379 ± 0.0003	6.81 ± 0.07	0.0460 ± 0.0003	8.27 ± 0.08
Dataset 2	0.0165 ± 0.0005	3.33 ± 0.12	0.0232 ± 0.0022	4.65 ± 0.48
Dataset 3	0.0327 ± 0.0021	6.68 ± 0.39	0.0434 ± 0.0008	8.76 ± 0.15

4.1.3. Gaussian Convolution Kernel

The Gaussian convolution kernel in the model discussed in this paper is replaced with a traditional convolution kernel. Random initialization is used for training. Figure 13 and Table 8 show the prediction results of the two models.

**Figure 13.** The prediction comparison diagram of Gaussian convolution kernel.**Table 8.** The prediction comparison table of differential processing of input variables.

	Control Group		Experimental Group	
	MAE	MAPE/%	MAE	MAPE/%
Dataset 1	0.0379 ± 0.0003	6.81 ± 0.07	0.0440 ± 0.0006	7.90 ± 0.11
Dataset 2	0.0165 ± 0.0005	3.33 ± 0.12	0.0185 ± 0.0001	3.74 ± 0.02
Dataset 3	0.0327 ± 0.0021	6.68 ± 0.39	0.0398 ± 0.0006	8.04 ± 0.17

From the results, it can be seen that the model in this paper achieves a better prediction performance. The convolution operation in this paper extracts inertial and delay information from the data by employing the Gaussian kernel. There are two trainable parameters in the convolution kernel used in this paper, which determine the shape of the Gaussian curve. Therefore, the convolution kernels used in this paper can learn well for different inertia and delay features. The analysis is carried out in two aspects.

(1) The problem of data characteristics: Due to the large delay and inertia characteristics of CFB unit data, it is difficult for the traditional CNN algorithm to identify the inertia and delay characteristics of each feature. Traditional CNN is used to mine periodic information in a time series. In contrast, the bed pressure prediction task has a short time interval and is time sensitive. Therefore, the periodic information and periodic patterns of the data are not obvious in this task for CFB units.

(2) The problem of data size: The model requires more inductive bias to improve prediction performance on the small datasets [39]. The Gaussian convolution kernel in this model contains more inductive bias. Therefore, CNN with Gaussian convolutional kernels performs better on smaller datasets than traditional CNN.

4.1.4. Gaussian Convolution Kernel

Ablation experiments were performed for segmentation training. The experimental results are shown in Figure 14 and Table 9. From the results, it can be seen that the model produces better prediction performance after using segmented training. The main reason for this is related to the attentional mechanism of TPA-LSTM. In the original TPA-LSTM model, the model uses the last-moment information from the output of the LSTM model as the query vector in the attention mechanism. The output of the historical moment is attended to using the query vector. However, in the directly trained model, the last-moment information greatly differs from the target information. Therefore, during the training of the attention layer, the attention information of the historical moment also has a large deviation. In the segmented training, the parameters of the LSTM layer are obtained by the first segment training. At the end of the first training segment, the last moment of information in the LSTM output captures the target information very accurately. Subsequent training of the parameters of the attention layer resulted in faster convergence of the model, and better prediction performance was achieved.

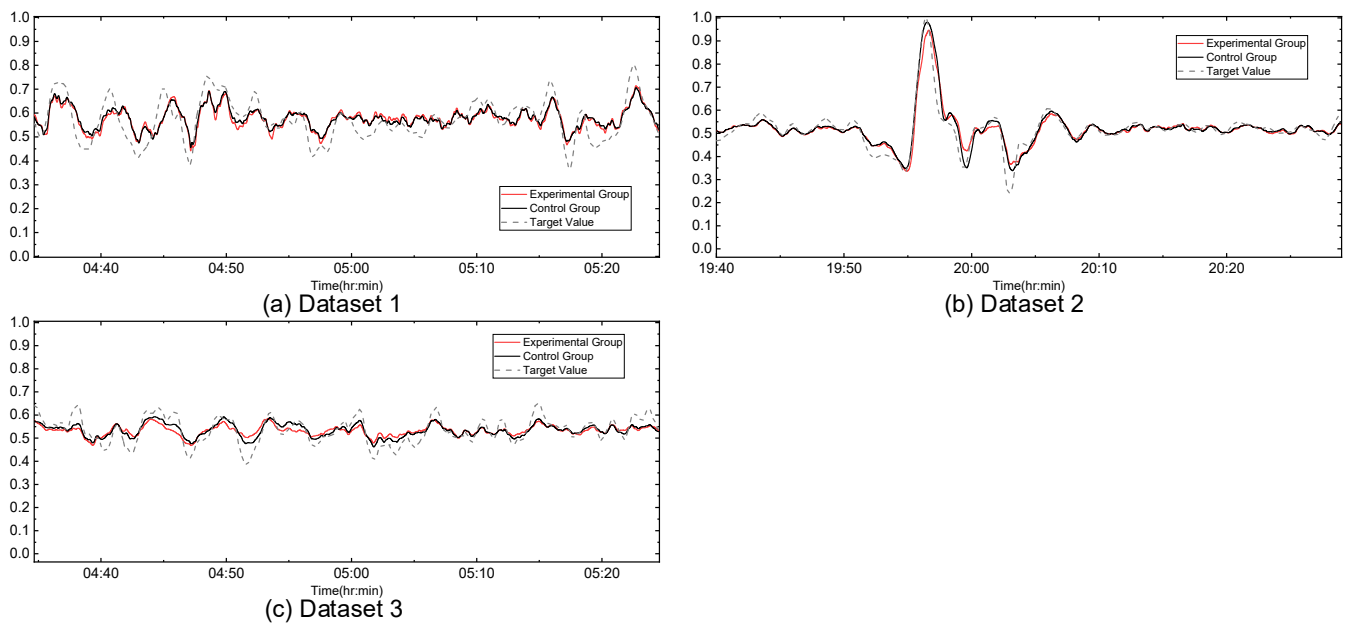


Figure 14. The prediction comparison diagram of segmented training.

Table 9. The prediction comparison table of segmented training.

	Control Group		Experimental Group	
	MAE	MAPE/%	MAE	MAPE/%
Dataset 1	0.0379 ± 0.0003	6.81 ± 0.07	0.0401 ± 0.0004	7.17 ± 0.14
Dataset 2	0.0165 ± 0.0005	3.33 ± 0.12	0.0178 ± 0.0005	3.59 ± 0.07
Dataset 3	0.0327 ± 0.0021	6.68 ± 0.39	0.0384 ± 0.0011	7.75 ± 0.21

4.2. Comparison of Prediction Models

The CNN-LSTM [40], the least squares support vector regression (LSSVR), LSTM [33], BP neural network, and the TPA-LSTM [26] model are compared to verify the validity of the model examined in this paper. In the comparison experiments discussed in this section, the input variables of the model are the six variables screened in Section 3.2 and

their differential values. The output of the model is the differential value of bed pressure. Figures 15 and 16, and Table 10 show the comparison results.

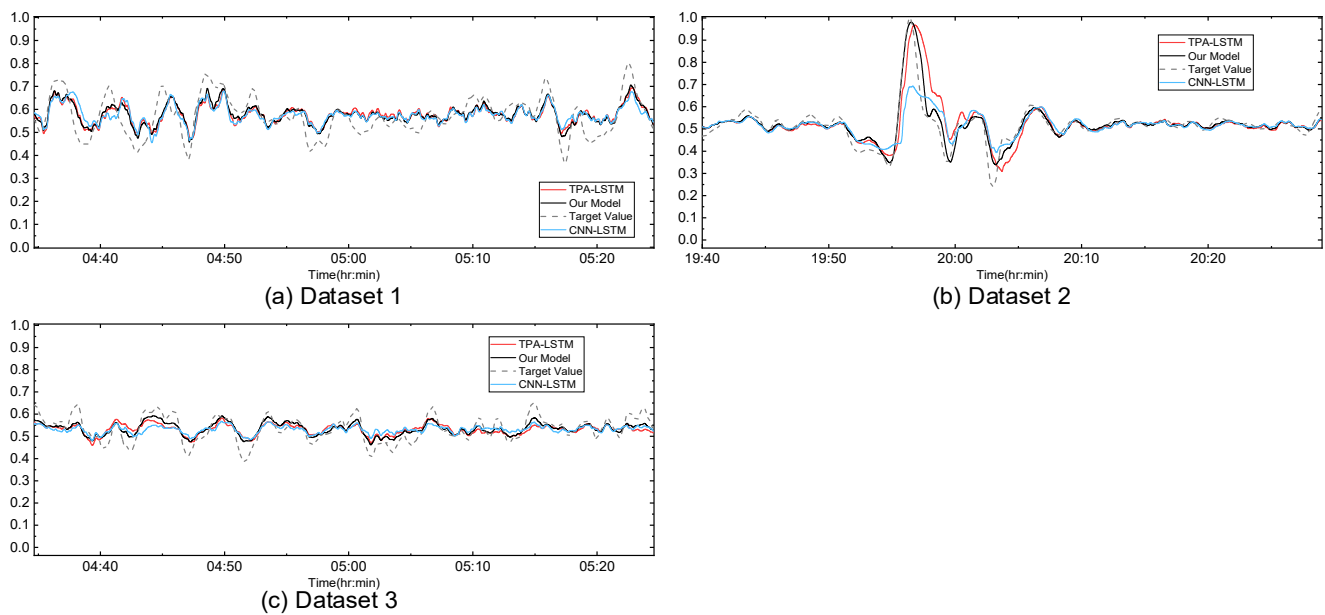


Figure 15. The prediction comparison diagram of differential value.

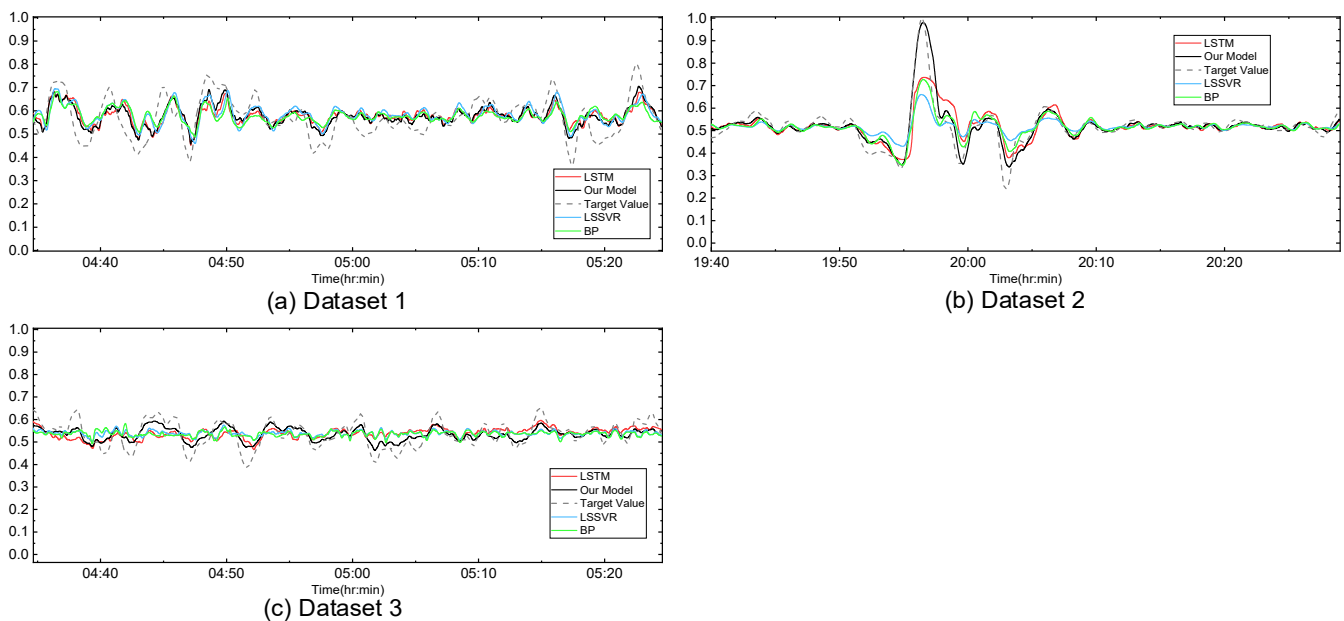


Figure 16. The prediction comparison diagram of bed pressure.

Table 10. Comparison results table.

	Dataset 1		Dataset 2		Dataset 3	
	MAE	MAPE/%	MAE	MAPE/%	MAE	MAPE/%
TPA-LSTM	0.0433 ± 0.0004	7.75 ± 0.07	0.0220 ± 0.0006	4.38 ± 0.14	0.0412 ± 0.0017	8.27 ± 0.30
CNN-LSTM	0.0454 ± 0.0003	8.09 ± 0.04	0.0223 ± 0.0011	4.38 ± 0.21	0.0418 ± 0.0026	8.46 ± 0.49
LSTM	0.0472 ± 0.0005	8.47 ± 0.14	0.0222 ± 0.0001	4.41 ± 0.06	0.0455 ± 0.0022	9.16 ± 0.40
LSSVR	0.0496	8.94	0.0251	5.03	0.0478	9.84
BP	0.0519 ± 0.0006	9.30 ± 0.18	0.0226 ± 0.0001	4.49 ± 0.03	0.0496 ± 0.0004	10.07 ± 0.02
Our model	0.0379 ± 0.0003	6.81 ± 0.07	0.0165 ± 0.0005	3.33 ± 0.12	0.0327 ± 0.0021	6.68 ± 0.39

From the results, it can be seen that the model examined in this paper achieves the best prediction performance on all three datasets. Dataset 1 and Dataset 3 are data from the regular operation of the unit. It can be seen from the figure that the model in this paper can better learn the characteristics of the bed pressure variation values. The test set of Dataset 2 contains fault data. Our model can also accurately predict the bed pressure variation in this case.

The BP model and the LSSVR model cannot handle the time series information of the data in the input of the model. Therefore, both models perform worse than the other algorithms in all three datasets. To our surprise, the BP model performs close to the other compared algorithms in Dataset 2. After analysis, it is revealed that the main reason for this is the autocorrelation of the data. The sampling time is too short, resulting in autocorrelation between the data so that the temporal information of the data is not fully displayed. The LSTM model performs the selection of historical states through the forgetting gate. However, the inertia and delay properties unique to each feature are not fully considered during the LSTM model computation. Therefore, the model does not learn the target values correctly. The convolution operations in the TPA-LSTM and CNN-LSTM models are used to extract the period information in the MTS data. This type of period information is not obvious during the short-term operation of CFB units. Therefore, the CNN-LSTM model and TPA-LSTM cannot learn the periodic information. This problem leads to the degradation of the prediction performance of the models.

5. Conclusions

This paper proposes a prediction method of the bed pressure of circulating fluidized bed units. The method is verified by the operation data of Linhuan Zhongli 1 # 330 mw circulating fluidized bed unit in China, which operates with coal slime blending. In terms of MAE metrics, the proposed method reached 0.0379 kPa, 0.0165 kPa, and 0.0327 kPa on the three datasets, respectively. In terms of MAPE metrics, the proposed method reached 6.81%, 3.33%, and 6.68 % on the three datasets, respectively. The results of this paper have particular reference significance for applying in-depth learning in the industrial field. The predictive method proposed in this paper has achieved an excellent prediction effect on actual operation data, which effectively proves the reliability of the prediction model. However, the limitation of this study is that the effects of coal quality variation and coal slime blending ratio on model accuracy are not considered. Future work includes the method needed to characterize coal quality data and coal slime blending ratios for the deep learning model.

Author Contributions: Conceptualization, J.C., F.H. and M.G.; methodology, J.C. and F.H.; software, J.C.; validation, J.C., F.H. and M.G.; formal analysis, J.C.; investigation, F.H.; resources, M.G. and F.H.; data curation, J.C. and F.H.; writing—original draft preparation, J.C.; writing—review and editing, F.H.; visualization, J.C.; supervision, F.H.; project administration, F.H.; funding acquisition, M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 52006062) and the Fundamental Research Funds for the Central Universities (2020MS013).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Khanpit, V.; Tajane, S.P.; Mandavgane, S.A. Experimental Studies on Coal-Water Slurry Fuel Prepared from Pretreated Low-Grade Coal. *Int. J. Coal Prep. Util.* **2022**, *42*, 831–845. [[CrossRef](#)]
2. Nowak, W. Clean Coal Fluidized-Bed Technology in Poland. *Appl. Energy* **2003**, *74*, 405–413. [[CrossRef](#)]

3. Li, J.J.; Hu, N.; Yao, X.; Yang, S. Experimental Study of the Bed Inventory Overturn in Pant-Legs Furnace of CFB Boiler. *Zhongguo Kuangye Daxue Xuebao/J. China Univ. Min. Technol.* **2011**, *40*, 54–59.
4. Sun, J.; Wang, Z.; Cao, W.; Wu, H.; Li, Z.; Lei, X.; Kuang, W. Mechanism of the Impact of Particle Size Distribution to Bed-Inventory Overturn for Pant-Leg Circulating Fluidized Bed. *Flow Turbul. Combust.* **2013**, *90*, 885–895. [[CrossRef](#)]
5. Wang, S.; Luo, K.; Hu, C.; Sun, L.; Fan, J. Effect of Superficial Gas Velocity on Solid Behaviors in a Full-Loop CFB. *Powder Technol.* **2018**, *333*, 91–105. [[CrossRef](#)]
6. Yang, C.; Duan, Y.; Hu, H. Application of CFD-DEM to the Study of Solid Exchange in a Dual-Leg Fluidized Bed. *Particuology* **2013**, *11*, 636–646. [[CrossRef](#)]
7. Wang, C.; Zhang, J.; Lan, X.; Gao, J.; Zhu, J. Quantitative Study of the Gas-Solids Flow and Its Heterogeneity/Nonuniformity in a 14 m Two-Dimensional CFB Riser Reactor. *Ind. Eng. Chem. Res.* **2020**, *59*, 437–449. [[CrossRef](#)]
8. Liu, H.; Li, J.; Wang, Q. Simulation of Gas–Solid Flow Characteristics in a Circulating Fluidized Bed Based on a Computational Particle Fluid Dynamics Model. *Powder Technol.* **2017**, *321*, 132–142. [[CrossRef](#)]
9. Gungor, A. Predicting Axial Pressure Profile of a CFB. *Chem. Eng. J.* **2008**, *140*, 448–456. [[CrossRef](#)]
10. Zhang, Y.; Zhang, M.; Zhu, S.; Huang, Y.; Deng, B.; Gao, X.; Jiang, X.; Lyu, J.; Yang, H. Mechanism Analysis of Gas Solid Flow Non-Uniformity Problem of 330 MW CFB Boiler. *Chem. Eng. Res. Des.* **2019**, *145*, 258–267. [[CrossRef](#)]
11. Pang, J.; Zhang, N.; Xiao, Q.; Qi, F.; Xue, X. A New Intelligent and Data-Driven Product Quality Control System of Industrial Valve Manufacturing Process in CPS. *Comput. Commun.* **2021**, *175*, 25–34. [[CrossRef](#)]
12. Adams, D.; Oh, D.H.; Kim, D.W.; Lee, C.H.; Oh, M. Deep Reinforcement Learning Optimization Framework for a Power Generation Plant Considering Performance and Environmental Issues. *J. Clean. Prod.* **2021**, *291*, 125915. [[CrossRef](#)]
13. Cui, Y.; Liu, H.; Wang, Q.; Zheng, Z.; Wang, H.; Yue, Z.; Ming, Z.; Wen, M.; Feng, L.; Yao, M. Investigation on the Ignition Delay Prediction Model of Multi-Component Surrogates Based on Back Propagation (BP) Neural Network. *Combust. Flame* **2022**, *237*, 111852. [[CrossRef](#)]
14. Adams, D.; Oh, D.H.; Kim, D.W.; Lee, C.H.; Oh, M. Prediction of SO_x–NO_x Emission from a Coal-Fired CFB Power Plant with Machine Learning: Plant Data Learned by Deep Neural Network and Least Square Support Vector Machine. *J. Clean. Prod.* **2020**, *270*, 122310. [[CrossRef](#)]
15. Li, Y.; Zhang, M.; Chen, C. A Deep-Learning Intelligent System Incorporating Data Augmentation for Short-Term Voltage Stability Assessment of Power Systems. *Appl. Energy* **2022**, *308*, 118347. [[CrossRef](#)]
16. Yu, H.; Gao, M.; Zhang, H.; Chen, Y. Dynamic Modeling for SO₂–NO_x Emission Concentration of Circulating Fluidized Bed Units Based on Quantum Genetic Algorithm—Extreme Learning Machine. *J. Clean. Prod.* **2021**, *324*, 129170. [[CrossRef](#)]
17. Xia, M.; Shao, H.; Ma, X.; De Silva, C.W. A Stacked GRU-RNN-Based Approach for Predicting Renewable Energy and Electricity Load for Smart Grid Operation. *IEEE Trans. Ind. Inform.* **2021**, *17*, 7050–7059. [[CrossRef](#)]
18. Shahid, F.; Zameer, A.; Muneeb, M. A Novel Genetic LSTM Model for Wind Power Forecast. *Energy* **2021**, *223*, 120069. [[CrossRef](#)]
19. Inapakurthi, R.K.; Miriyala, S.S.; Mitra, K. Deep Learning Based Dynamic Behavior Modelling and Prediction of Particulate Matter in Air. *Chem. Eng. J.* **2021**, *426*, 131221. [[CrossRef](#)]
20. Mnih, V.; Heess, N.; Graves, A.; Kavukcuoglu, K. Recurrent Models of Visual Attention. In Proceedings of the Twenty-Eighth Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Volume 3.
21. Bahdanau, D.; Cho, K.H.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015.
22. Xu, K.; Ba, J.L.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.S.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In Proceedings of the 32nd International Conference on Machine Learning (ICML 2015), Lille, France, 6–11 July 2015; Volume 3.
23. Zhang, Z.; Zou, Y.; Gan, C. Textual Sentiment Analysis via Three Different Attention Convolutional Neural Networks and Cross-Modality Consistent Regression. *Neurocomputing* **2018**, *275*, 1407–1415. [[CrossRef](#)]
24. Azam, M.F.; Younis, S. Multi-Horizon Electricity Load and Price Forecasting Using an Interpretable Multi-Head Self-Attention and EEMD-Based Framework. *IEEE Access* **2021**, *9*, 85918–85932. [[CrossRef](#)]
25. Xie, P.; Gao, M.; Zhang, H.; Niu, Y.; Wang, X. Dynamic Modeling for NO_x Emission Sequence Prediction of SCR System Outlet Based on Sequence to Sequence Long Short-Term Memory Network. *Energy* **2020**, *190*, 116482. [[CrossRef](#)]
26. Shih, S.Y.; Sun, F.K.; Lee, H.Y. Temporal Pattern Attention for Multivariate Time Series Forecasting. *Mach. Learn.* **2019**, *108*, 1421–1441. [[CrossRef](#)]
27. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 2017.
28. Reh, L. Development Potentials and Research Needs in Circulating Fluidized Bed Combustion. *China Particuol.* **2003**, *1*, 185–200. [[CrossRef](#)]
29. Li, J.; Zhang, H.; Yang, H.; Liu, Q.; Yue, G. The Mechanism of Lateral Solid Transfer in a CFB Riser with Pant-Leg Structure. *Energy Fuels* **2010**, *24*, 2628–2633. [[CrossRef](#)]
30. Smolders, K.; Baeyens, J. Elutriation of Fines from Gas Fluidized Beds: Mechanisms of Elutriation and Effect of Freeboard Geometry. *Powder Technol.* **1997**, *92*, 35–46. [[CrossRef](#)]

31. Ke, X.; Engblom, M.; Yang, H.; Brink, A.; Lyu, J.F.; Zhang, M.; Zhao, B. Prediction and Minimization of NO_x Emission in a Circulating Fluidized Bed Combustor: A Comprehensive Mathematical Model for CFB Combustion. *Fuel* **2022**, *309*, 122133. [[CrossRef](#)]
32. Zhou, X.; Niu, T.; Xin, Y.; Li, Y.; Yang, D. Experimental and Numerical Investigation on Heat Transfer in the Vertical Upward Flow Water Wall of a 660 MW Ultra-Supercritical CFB Boiler. *Appl. Therm. Eng.* **2021**, *188*, 116664. [[CrossRef](#)]
33. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)]
34. Zhang, H.; Gao, M.; Fan, H.; Zhang, K.; Zhang, J. A Dynamic Model for Supercritical Once-through Circulating Fluidized Bed Boiler-Turbine Units. *Energy* **2022**, *241*, 122914. [[CrossRef](#)]
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; Volume 2016. [[CrossRef](#)]
36. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning (ICML 2015), Lille, France, 6–11 July 2015; Volume 1.
37. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019.
38. Xu, Z.Q.J.; Zhang, Y.; Luo, T.; Xiao, Y.; Ma, Z. Frequency Principle: Fourier Analysis Sheds Light on Deep Neural Networks. *Commun. Comput. Phys.* **2020**, *28*, 1746–1767. [[CrossRef](#)]
39. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
40. Shipman, R.; Roberts, R.; Waldron, J.; Naylor, S.; Pinchin, J.; Rodrigues, L.; Gillott, M. We Got the Power: Predicting Available Capacity for Vehicle-to-Grid Services Using a Deep Recurrent Neural Network. *Energy* **2021**, *221*, 119813. [[CrossRef](#)]